DOCUMENT RESUME

ED 204 398 TM 810 398

AUTHOR Arneklev, Bruce L.
TITLE Scaling Objective-Based Curriculum.
PUB DATE 17 Apr 81
NOTE 26p.: Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981). Small type in tables.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement: *Behavioral Objectives: *Criterion Referenced Tests: Elementary Education: Elementary School Mathematics: *Item Banks: Language Tests: *Latent Trait Theory: Quantitative Tests: Reading Tests: *Scaling
IDENTIFIERS Comprehensive Tests of Basic Skills: *Rasch Model: *Test Linking

ABSTRACT

Objective referenced tests were scaled using the Rasch latent trait model to measure Student Learning Objectives (SLOs) in grades one through eight. The tests, in the areas of reading, language and math, were scaled and linked through students, blocks of items, and test item banks. The study tests latent trait theory in a public school setting. General assumptions are substantiated. Limitations that should be considered by developers and users of test item banks are clarified, and some questions are raised about the appropriateness of expecting large increments of measured growth among students, particularly at higher grade levels. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED204398

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

SCALING OBJECTIVE-BASED CURRICULUM

by

Bruce L. Arnekiev

Office of Research and Evaluation

Tacoma Public Schools

P.O. Box 1357

Tacoma, Washington 98401

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. L. Arnekiev

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 810 398

Paper prepared for presentation at the

American Educational Research Association annual meeting

in Los Angeles, California, April 17, 1981

ABSTRACT

Objective referenced tests were scaled using the Rasch latent trait model. Eight grade level related tests in each of the areas of reading, language, and math were scaled and linked through students, blocks of items, and test item banks.

The study tests latent trait theory in a public school setting. General assumptions are substantiated. Limitations that should be considered by developers and users of test item banks are clarified, and some questions are raised about the appropriateness of expecting large increments of measured growth among students, particularly at higher grade levels.

SCALING OBJECTIVE-BASED CURRICULUM *

The instruments used to quantify units of physical phenomena (e.g., height, weight, temperature, etc.) are taken for granted as possessing the power to describe reality in stable increments. That confidence arises in part because results from different instruments used in the physical sciences can readily be compared. On the other hand, in education much difficulty has been encountered in attempts to communicate how much learning has occurred. Reporting about "muchness" of learning has been meaningful only in terms of particular tests and in relation to particular reference or norm groups. The advent of latent trait logistic models (Baker 1977) will hopefully assist educational evaluators to make their efforts more fruitful. This paper describes some outcomes from one attempt to apply the Rasch latent trait model in a public school setting.

A few authors (Wright 1977; Forster and Doherty 1978; and Holmes 1978) have reported encouraging results using Rasch scaling methodology. On the other hand, Slinde and Linn (1977), Whitely (1977), Plake and Hoover (1979), Bejar (1980), Kolen (1980), Loyd and Hoover (1980), and Yen (1980) have been more reserved in their enthusiasm for the one parameter model. Generally, enthusiasm has been greater for "horizontal" than "vertical" scaling. Horizontal means scaling of tests or items of similar difficulty and related to similar content. Vertical means scaling tests or items which were designed to measure different levels of ability and varied content within specified domains.

---

*The author gratefully acknowledges the contribution of Leonard T. Winchell of the Tacoma Public Schools for the aggregation of the data in Table VII, and for editorial contributions to this paper.

This paper will focus on results from scaling tests which were designed to measure Student Learning Objectives (SLOs) in grades one through eight in reading, language, and math. These objective referenced tests were designed (Arneklev 1979) primarily by using classical correlational techniques to evaluate the quality of items, sub-tests and total tests. After development Rasch scaling (Forster 1978) was applied.

PROCEDURE

The items on the twenty-four tests were scaled (1) through students to three levels of a norm-referenced test battery, (2) through blocks of items common to different levels within each content area, and (3) to a test item bank for each content area. The initial linking patterns were established on the basis of post-test data collected from approximately two thousand students in each grade level during the spring of 1979. At that time appropriate levels of the Comprehensive Tests of Basic Skills (CTBS) Form S were administered to substantial numbers of the same students within two weeks of post-testing with grade level SLO tests.

The scale unit used in all analysis and reporting was the "Rit", meaning Rasch Unit. This is a convenient increment adopted by the Northwest Evaluation Association (NWEA) to transform the basic logit to a type of standard score centered at two hundred (early fifth grade achievement) with a deviation of ten Rits per logit. Use of Rit values serves to eliminate much of the need to work with negative numbers and decimals associated with logistic values.

Insert Tables I, II, and III about here

The data and linking patterns reflected on Tables I, II, and III serve as an empirical base for deriving the equivalency of raw scores on different tests within each content area (Rentz and Bashaw 1977). It may be noted that values attained in scaling do not always end up in a "zero sum game," as would be desired in a perfect linking pattern. "Judgement" was used to maximize the symmetry of each linking net. In this process the scientist may become uneasy as some artistic manipulation of linking values is undertaken on the basis of the types and numbers of items in each link. Judgement must be exercised as to where "adjustments" must be made to bring the overall net into balance. In this study data from norm-referenced tests were weighted heavily, because of their content (more general) and their length (85, 98, or 105 items).

Linking values attained generally averaged algebraically to within two Rits; however, there were serious shortcomings in the data (or model) at grade seven in reading. The development of higher mental processes at this age may be one of the confounding factors. Whatever the cause, further study of this area is of great importance. Fortunately the availability of alternate links in conjunction with projection of values in terms of a monotonic learning growth curve served to provide a best guess of which values were most likely to be correct. Given all possible pathways, the overall net within each content area provides a good starting point from which to line up scores in a validation process.

Insert Tables IV, V, and VI about here

RESULTS

A review of concurrence between individual student scores was made as a preliminary check of linking values. Data from one classroom are depicted

in Table IV. A more illustrative presentation of the data is made in Tables V and VI for respective content areas. From this data it becomes apparent that:

1. Linking values are derived from group data. Even though latent trait models are advertized as sample independent, the scaling process serves to average out individual differences.
2. There is a serious ceiling effect. Scores above ninety percent correct do not predict well.
3. Low scores often do not predict well. Examination of individual protocols revealed that scores below forty percent correct were frequently caused by the failure of a student to attempt or complete certain sub tests.
4. The shape of the test score distribution is very important. Table VI depicts data from a relatively normal distribution on the SLO test predicting to the high end of the CTBS (Level 1 of the CTBS math test is relatively easy for students at the end of fourth grade). The specificity of the SLO test also tends to restrict the range of achievement measured in contrast to the more general content on the CTBS.

These four considerations and the general appearance of Tables V and VI serve to reinforce appreciation for error of measurement.

Insert Table VII about here

A more complete discriptive analysis of score concurrence was accomplished by way of data processing with the Statistical Package for the Social Sciences (SPSS). Those data are provided in Table VII. A median correlation of .80 was attained between Rit scores attained through concurrent testing with SLO and

CTBS measures within each content area. This moderately high indication of concurrence could be improved by correction for attenuation due to error of measurement, but what is probably most critical for one wanting to compare accuracy of horizontal scaling is found in difference between means.

Only six of the twenty-one pairs of mean Rit scores depicted on Table VII were found to be equivalent at the .05 level of confidence. In interpreting these differences it should be noted that all linking values were computed by hand and rounded to the nearest Rit (.1 logit). Computer technology could contribute to greater accuracy and reliability. However, the nine Rit difference between second grade SLO vs CTBS math scores is clearly a real difference.

One may agree that the curriculum domains mapped by SLO and CTBS tests at the second grade level is sufficiently different to prevent concurrence. Indeed the Level 1 CTBS covers multiplication and division which are not generally taught to second graders, and therefore those kinds of items are not to be found on the SLO test. This incongruence suggests that scaling must be restricted to predicting horizontal equivalence only when content concurs very closely.

In spite of the inconsistencies found in trying to establish exacting indications of concurrence, the overall scaling procedure has provided some very useful outcomes.

Insert Tables VIII, IX, and X about here

Tables VIII, IX, and X are now being used in the district to identify Title I and other students in need of remedial assistance. Scores less than forty percent correct or greater than ninety percent correct have been shaded on tables to indicate test ranges where measurement error is greatest. Pre- and post test averages for each grade level are plotted along with an estimated

grade equivalence on the right hand margin. This provides a vertical scale in terms of which to set appropriate expectations for growth, as well as to suggest the kind of curriculum or objectives where a student might function most efficiently.

These tables are a graphic representation of a desired outcome from vertical scaling. They provide a mechanism for estimating the equivalence of high score on an easy test to a lower score on a test that is more difficult. Average post-test Rit scores for the end of one year tend to articulate quite well with average pretest Rit scores at the beginning of the next higher grade. The only large discrepancy is between the end of grade one and the beginning of grade two in math. This could be explained by the proposition that students may forget relatively unused math skills over the summer. The general trend is for less "forgetting" with advancing grade levels. Of course this is average data and serves to mask individual differences.

The data serve to document a commonly held assumption that measured achievement gains are largest at lower grades. This is evident both in the decreasing magnitude of pre-post change with advancing grade levels, and in "estimated grade equivalents" derived on the basis of post-test scores in a district that is slightly above the national average on national norms. Perhaps the most profound finding of the study is the limited amount of measured growth that occurs in one year of instruction at any level after grades one and two.

After grade two the average amount of progression during a year of instruction appears to be less than ten percent of the items on any test, whether that test is norm or objective referenced. Certain students may demonstrate more growth, but forgetting and less progress by others in a

class serve to make the data less reinforcing for instructional effort than might be expected. This is most distressing when post-test scores turn out equal to or lower than pretest scores, after a full year of instruction for many individual students.

CONCLUSIONS

This endeavor was began in hopes of creating a curriculum scale what would enhance evaluation according to three postulates:

1. That normative information (N.C.E.s, National Percentiles, or Grade Equivalents) could be reported from locally developed tests. This is now being done for Title I and remedial assistance programs.
2. That results from tests matched to student achievement level would provide more adequate assessment within groups that contain retarded and/or advanced students. This has been attempted for small groups, but complications arising from all the irregularities injected into the scoring system have precluded districtwide implimentation to date. Questions about the robustness of the model for vertical scaling raise some doubts, but accuracy of measurement attained by functional level testing in conjunction with awareness of the shortcomings of other currently available alternatives make it a desired practice.
3. That new versions of locally developed or norm referenced tests can be adopted without losing baseline reference points for plotting achievement trends. This has been done with some confidence through horizontal scaling. Questions still exist as to where horizontal scaling leaves off, and about what limits must be imposed for vertical scaling.

A special appreciation has been gained for the need for more than one linkage in establsihing the difference of difficulty between two or more tests. Difficulties of items in banks are not as stable as theory or advertising might

lead the uninitiated to believe. Sensitivity to instruction and/or interactions between instruction and learning styles can alter "latent" difficulties substantially. Linkages established on data from grades four to six appear to be most stable. Before and after those years specificity of instruction and attitudes among teachers and students tend to increase errors of measurement and prediction. The liabilities of this new approach to measurement are serious, but potential rewards justify continued exploration.

Growing availability of item banks increases the opportunity to tailor tests to local curriculum, and to insure more accurate assessment by building tests with appropriate difficulty for specified individuals or groups. The opportunity to capitalize on these possibilities also increases the responsibilities for curriculum and test builders to treat tests as tools with recognized shortcomings and strengths. Continued scrutiny of efforts in linking tests is extremely important, because the procedures used by evaluators can significantly change reportable results.

REFERENCES

ARNEKLEV, BRUCE L. Tacoma Public Schools Learning Objectives Program. Paper presented at the American Educational Research Association annual meeting, San Francisco, California, April 9, 1979.

BAKER, FRANK B. Advances in item analysis. Review of Educational Research, Winter 1977, Vol. 47 #1, Pp 151-178.

BEJAR, ISSAC I. A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, Winter 1980, Vol. 17 #4, Pp 283-296

FORSTER, FRED The mathematics of the Rasch model. Northwest Evaluation Association Journal, December 1978, Vol 1 #1, Pp 23-38.
(ERIC Designation = ED 189 101)

FORSTER, F. & DOHERTY, V. Using the Rasch Model to Improve a District's Testing Program. A paper presented at the meeting of the American Educational Research Association, Toronto, Canada, March 1978.

HOLMES, SUSAN E. ESEA Title I Linking Project Final Report. Oregon Department of Education, 700 Pringle Parkway S.E., Salem, Oregon 97310, 1980.

KOLEN, MICHAEL S. Comparison of Traditional and Latent Trait Theory Methods for Equating Tests. Paper presented at the annual meeting of the American Educational Research Association in Boston, Massachusettes, April 1980.

LOYD, BRENDA A. & HOOVER, H.D. Vertical equating using the Rasch model. Journal of Educational Measurement, Fall 1980, Vol. 17 #3, Pp 179-193.

PLAKE, BARBARA S. & HOOVER, H.D. The comparability of equal raw scores obtained from in-level and out-of-level testing: One source of discrepancy between in-level and out-of-level grade equivalent scores. Journal of Educational Measurement, Fall 1979, Vol. 16 #3, Pp 217-278.

RENTZ, ROBERT, & BASHAW, W.L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, Summer 1977, Vol. 14 #2, Pp 161-179

SLINDE, JEFFREY A. & LINN, ROBERT L. Vertical equating: fact or phantom. Journal of Educational Measurement, Spring 1977, Vol. 14 #1, Pp 23-32.

WHITELY, SUSAN E. Models, meanings, and misunderstandings: Some issues in applying Rasch's theory. Journal of Educational Measurement, Fall 1977, Vol. 14 #3, Pp 227-235.

WRIGHT, B.D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton, N.J., Educational Testing Service, 1968, Pp 85-101.

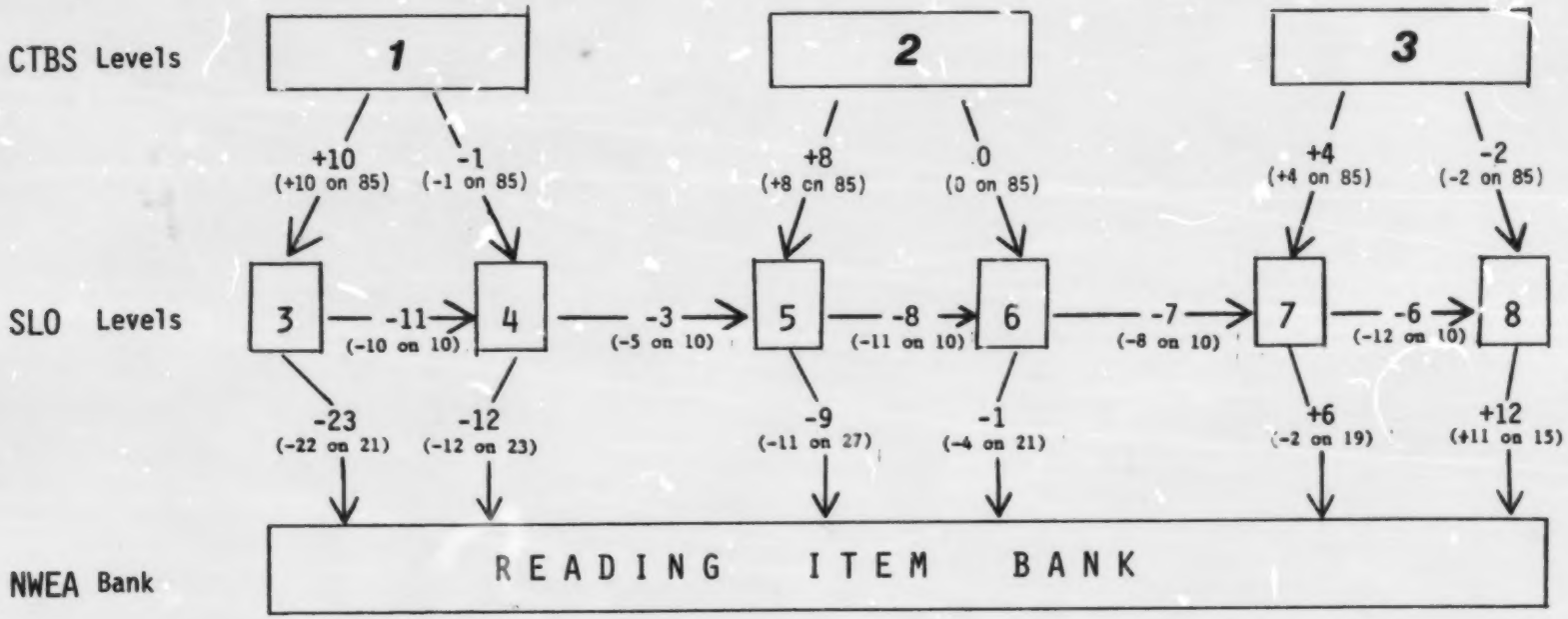
WRIGHT, BENJAMIN D Solving measurement problems with the Rasch model. Journal of Educational Measurement, Summer 1977, Vol. 14 #2, Pp 97-116.

YEN, WENDY M. The extent, causes, and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, Winter 1980, Vol. 17 #4, Pp 297-312.

TABLE I

READING LINKING NET*

(actual values attained with number of items in common)

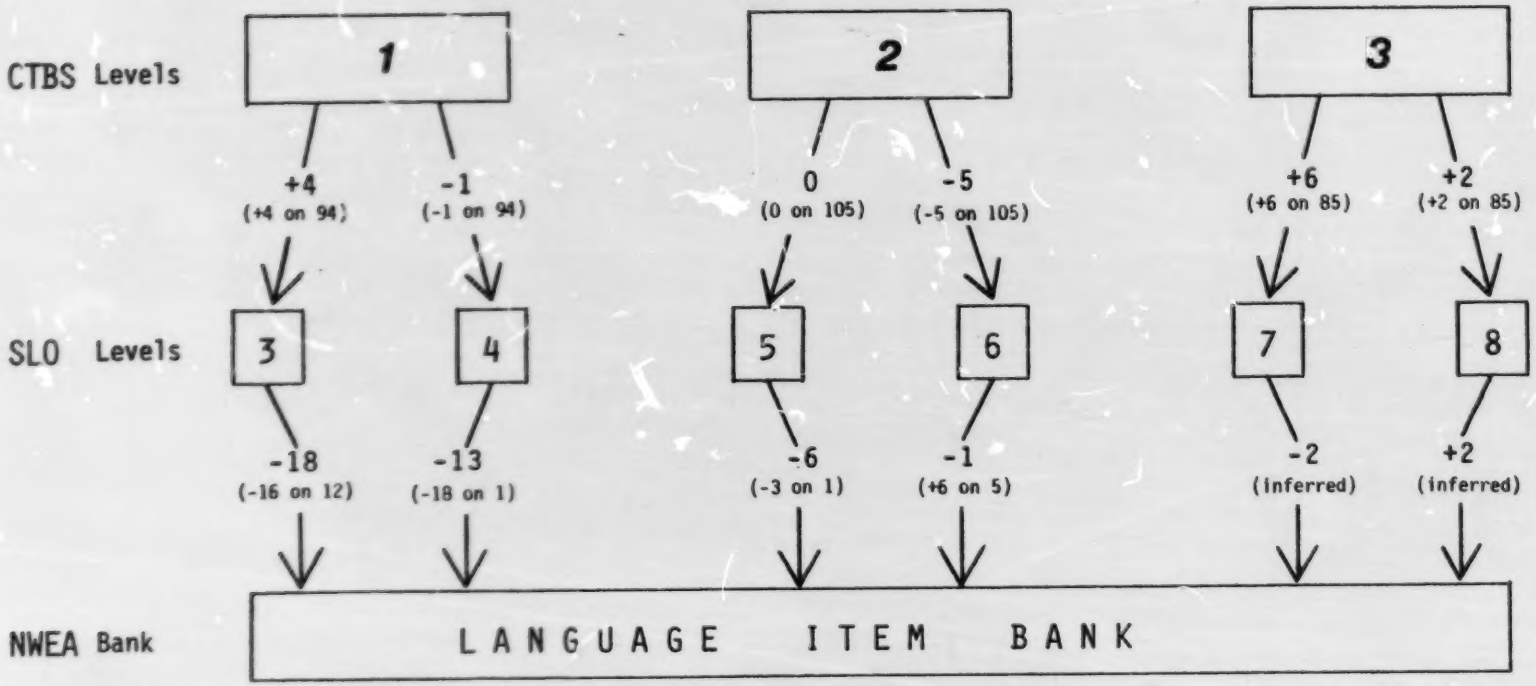


*Rit values used to link tests are printed in large type between boxes. Actual values attained and number of items from which they are attained are printed parenthetically under the values used. Attained values were adjusted to put total net into symmetry.

TABLE II

LANGUAGE LINKING NET*

(actual values attained with number of items in common)

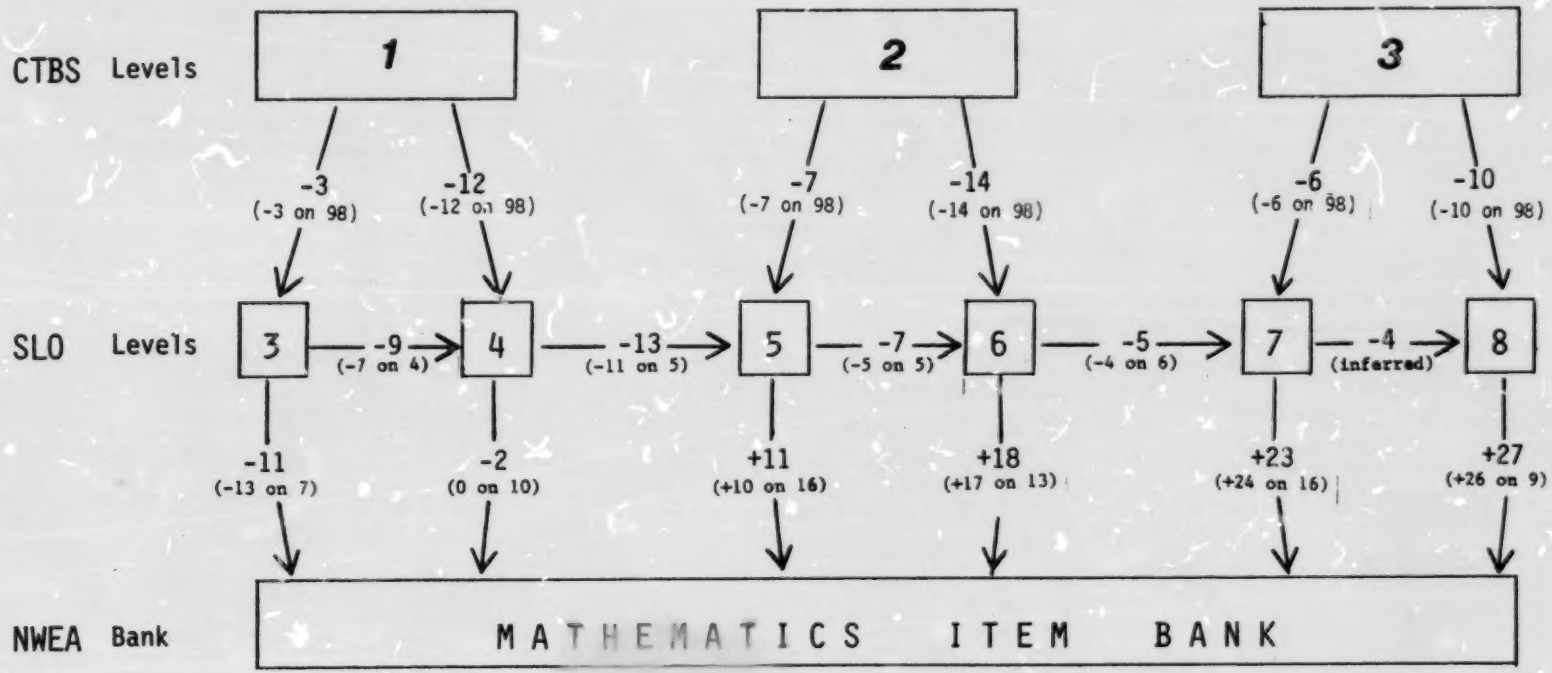


*Rit values used to link tests are printed in large type between boxes. Actual values attained and number of items from which they are attained are printed parenthetically under the values used. Attained values were adjusted to put total net into symmetry.

TABLE III

MATHEMATICS LINKING NET*

(actual values attained with number of items in common)



*Rit values used to link tests are printed in large type between boxes. Actual values attained and number of items from which they are attained are printed parenthetically under the values used. Attained values were adjusted to put total net into symmetry.

TABLE IV

TEST SCORES ATTAINED CONCURRENTLY

Student Name	READING				MATHEMATICS			
	Raw Score		Rit Score		Raw Score		Rit Score	
	SLO	CTBS	SLO	CTBS	SLO	CTBS	SLO	CTBS
A, Else	69	80	216	220	64	96	228	231
B, Janet	22	27	178	177	19	42	188	185
B, Deborah	67	71	212	207	47	75	210	202
C, Diane	67	78	212	216	52	96	214	231
C, Shannon	64	72	208	208	41	83	205	208
D, Peggysue	23	30	179	179	23	43	192	185
E. John	60	52	204	193	52	95	214	227
E, David	64	65	208	202	59	94	221	224
E, Christopher	72	80	223	220	57	94	219	224
E, Leonard	54	66	199	202	43	95	207	227
E, Evonne	75	82	245	220	51	96	213	231
E, Deborah	71	83	220	230	62	94	225	224
F, Andrea	58	81	214	222	55	93	217	221
F, James	60	64	204	201	48	83	211	221
G, Kandi	50	58	196	197	52	83	214	208
G, Jeffery	63	55	207	195	54	91	216	217
G, Edward	52	67	198	203	44	78	208	204
H, Andrea	32	27	185	177	27	50	195	189
K, Teresa	68	67	214	203	46	82	209	207
K, Marie	69	75	216	211	56	97	218	238
K, Paul	18	26	175	177	19	27	188	177
M, Kane	37	43	188	187	36	54	202	191
M, Ross	63	69	207	205	43	69	207	214
M, Sharalee	60	79	204	218	42	76	206	202
M, Margeret	26	33	181	181	36	71	202	200
M, Christine	44	40	193	186	31	70	198	199
M, Marian	62	50	206	191	37	82	202	207
Correlation	.92		.89		.90		.87	
Mean	54.8	60.0	203.4	201.3	44.3	78.9	208.5	210.9
Standard Deviation	17.2	19.2	15.8	15.9	12.5	19.5	10.4	16.5

TABLE V

READING SCORES ATTAINED CONCURRENTLY BY INDIVIDUALS

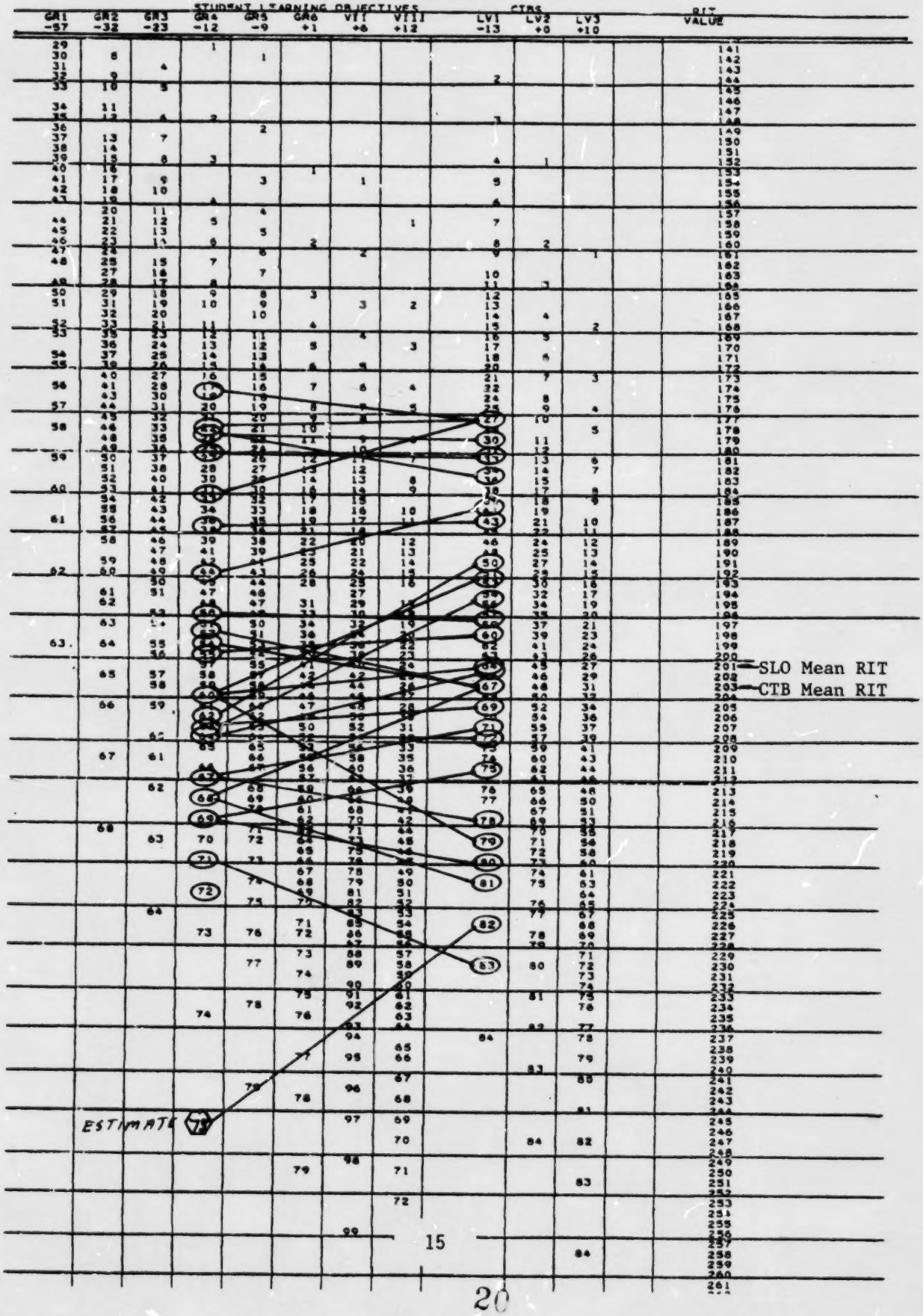


TABLE VI

MATHEMATICS SCORES ATTAINED CONCURRENTLY BY INDIVIDUALS

GR1 -52	GR2 -37	GR3 -0	STUDENT LEARNING OBJECTIVES GR4 +0	GR5 +10	GR6 +18	GR7 +23	GR8 +28	CTBS LV1 -12	LV2 +3	LV3 +17	RIT VALUE
38	19										139
40	20										140
42	22										141
43	23										142
45	25										143
47	26	1									144
49	28							2			145
51	30		1								146
53	31										147
55	33										148
57	35							3			149
59	37										150
61	39										151
63	41	2						4	1		152
65	43										153
66	45		2								154
68	47							5			155
70	50	3									156
72	52							6			157
74	54							7			158
75	56		3						2		159
77	59	4						8			160
79	61							9			161
80	63										162
82	65	5	4					10			163
83	68			1				11	3		164
85	70	6						12			165
86	72							13			166
87	73	7	5					14	4		167
89	77							15			168
90	79	8	6		1			16	5	1	169
91	81							18			170
92	84	9	7	2				19	6		171
93	86	10	7					20			172
94	88							22	7		173
95	90	11	8			1		23			174
96	92	12	9					25	8		175
97	94	13		3	2			28	9	2	176
98	97		10					29	10		177
99	98	14						30			178
100	100	15	11	4				32	11		179
	101	16	12				1	34	12	3	180
101	103	17	13	5	3	2		35	13		181
102	104	18						37	14		182
	106	19	14					39	15	4	183
103	107	20	15	6	4			41	17		184
	109	21	16	7				43	18	5	185
104	110	22	17		5	3		45	19		186
	111	24	18	8				47	21	6	187
	113	25	19				2		22		188
105	114	26	20	9	6	4		51	24	7	189
	115	27	21	10	7			53	25		190
	116	28	22	11				55	27	8	191
106	117	29	23		8	5	3	57	28	9	192
	118	31	24	12				59	30	10	193
	119	32	25	13	9	6		61	32	11	194
	121	33	27	14	10		4	63	34	12	195
107	120	34	28	15	11	7		65	36	13	196
	121	36	29	16	12			67	37	14	197
	122	37	31	17	13	8	5	69	39	15	198
	123	38	32	18	14	9		70	41	16	199
		39	34	20	15		6	72	43	18	200
		40	35	21	16	10		74	45	19	201
	124	41	36	22	17	11	7	76	47	20	202
		42	38	23	18	12		78	49	22	203
108	125	43	39	25	19		8	80	51	24	204
		44	41	26	20	13	9	82	53	25	205
		45	42	27	21	14		84	55	27	206
	126	46	43	29	22	15	10	86	57	29	207
		47	45	30	24	16	11	88	59	31	208
		48	46	31	26	18	12	90	61	33	209
	127	49	47	32	28	19	13	92	63	35	210
		50	49	34	29	20	14	94	65	37	211
			50	35	31	21	15	96	67	39	212
		51	51	37	32	23	16	98	68	41	213
	128	52	52	38	34	24	17	100	70	43	214
			54	39	35	25	18	102	72	45	215
		53	55	41	37	27	19	104	73	47	216
			56	42	39	28	20	106	75	50	217
		54	58	44	40	30	22	108	76	52	218
			59	44	41	31	23	110	78	54	219
		55	61	46	43	33	24	112	79	56	220
			62	47	45	34	25	114	80	58	221
	129		64	49	47	36	27	116	82	60	222
		56	65	49	48	37	28	118	83	62	223
			67	50	50	39	29	120	84	64	224
			62	51	51	40	31	122	85	66	225
		57	69	52	53	42	32	124	86	68	226
			71	53	54	43	34	126	87	70	227
			64	53	56	45	35	128	88	72	228
			74	54	57	46	37	130		73	229
		58	76	55	58	48	38	132	89	75	230
			78	56	60	49	39	134	90	77	231
					61	51	41			78	232
			66	57	62	52	42		91	79	233
				58	63	53	43		92	81	234
					64	55	45			82	235
			67	59	65	56	46		93	83	236
		59			66	57	47			84	237
				60	67	58	49			85	238
					68	59	50		94	86	239
					69	60	51			87	240
			68	61	70	61	52			88	241
					71	62	53		95	89	242
						63	54			90	243
				62	72	64	55				244
						65	56			91	245
					73	66	57		96		246
					74		58			92	247
			69			67	59				248
				63	75	68	60			93	249
											250
						69	61			94	251
					78		62				252
						70			97		253
										95	254
					7						255
				64		71					256
							63			96	257
											258
						72					259
											260

SLO Mean RIT

CTB Mean RIT

TABLE VII

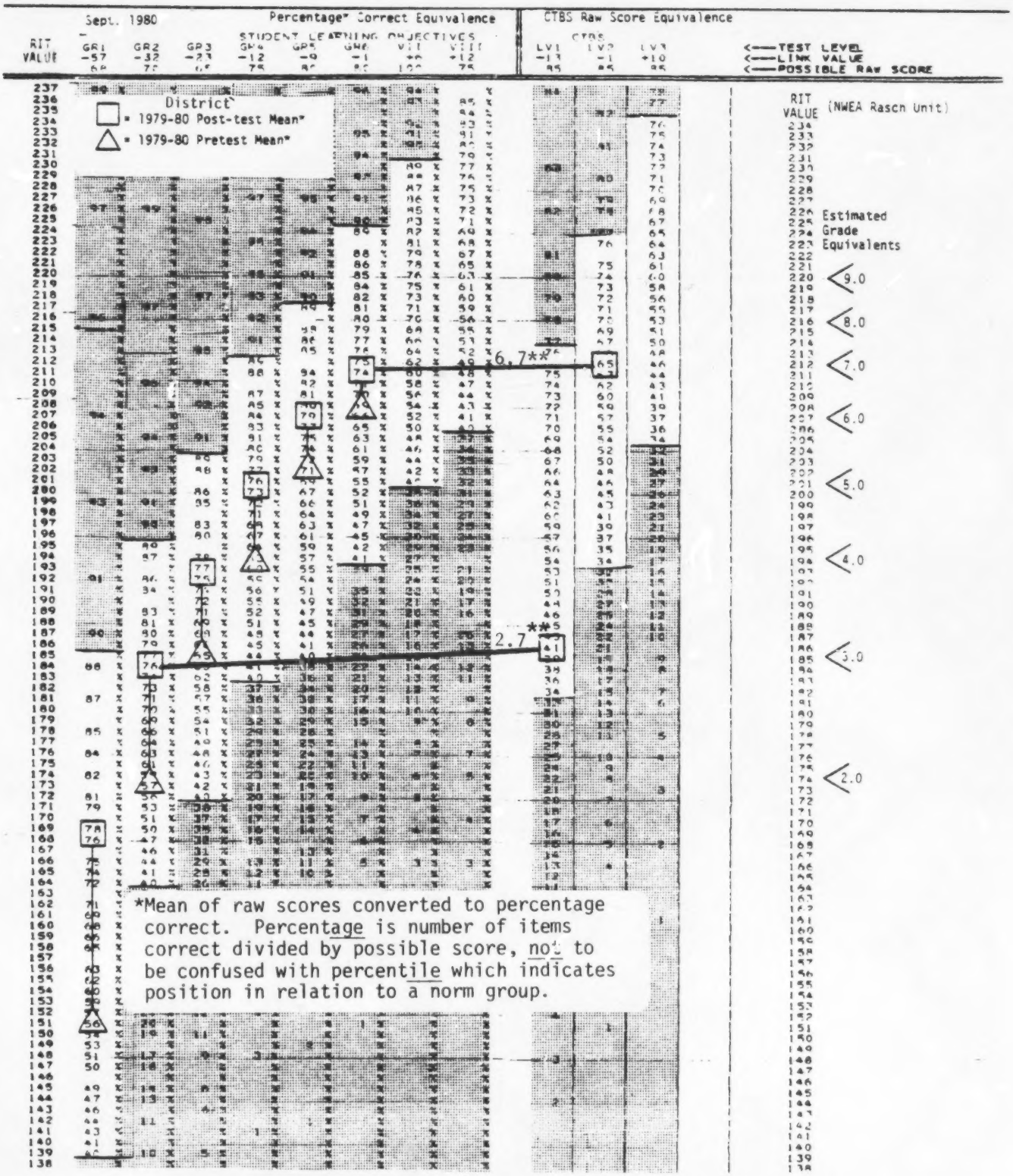
ACHIEVEMENT LEVEL CONCURRENCE ON CTBS AND SLO TESTS

Norm referenced CTBS test level	← I →									← II →						← III →					
Objective referenced SLO grade level	2			3			4			5			6			7			8		
Content area	Read	Lang	Math	Read	Lang	Math	Read	Lang	Math	Read	Lang	Math	Read	Lang	Math	Read	Lang	Math	Read	Lang	Math
Number of Students	1812	1232	1444	1161	1013	1209	1129	1065	1215	1273	1164	1368	1960	1948	1950	444	818	841	243	726	722
Mean Rit CTBS	184	188	183	194	195	198	202	202	207	207	207	214	213	211	221	216	213	225	224	217	230
Mean Rit SLO	188	190	192	194	197	198	201	201	206	209	207	214	214	210	219	219	214	225	225	222	230
CTBS Standard Deviation	11	8	10	14	11	13	15	12	14	11	10	11	12	10	11	11	9	12	10	10	13
SLO Standard Deviation	15	10	13	15	13	12	15	13	11	15	11	12	13	10	10	12	12	9	9	13	10
Pearson Correlation *(r)* not corrected for attenuation	.79	.64	.65	.81	.80	.81	.84	.80	.81	.80	.78	.82	.81	.78	.83	.77	.71	.83	.76	.71	.85

BLA/le
1/81

TABLE VIII

VERTICAL LINKING OF READING DATA

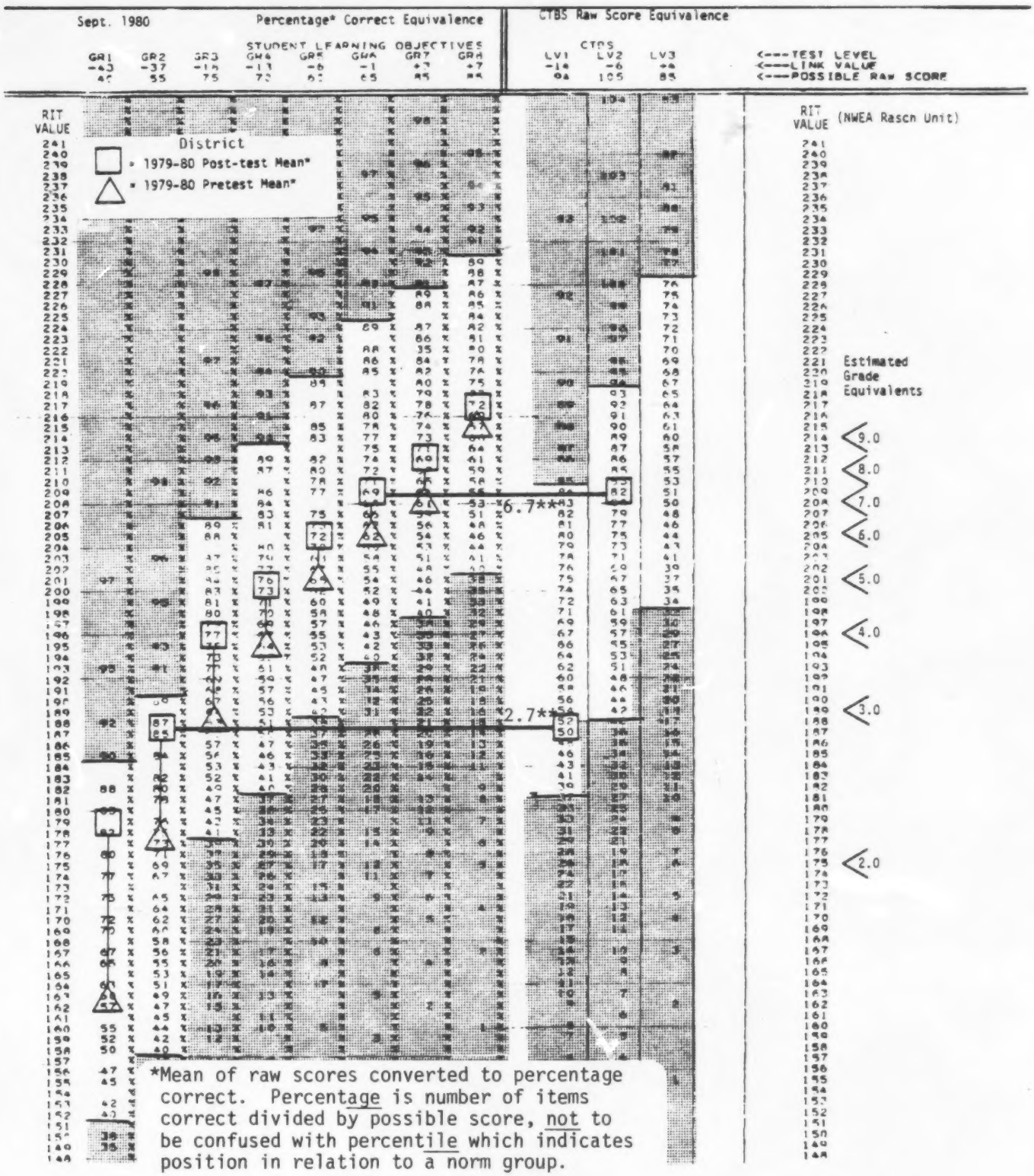


*Mean of raw scores converted to percentage correct. Percentage is number of items correct divided by possible score, not to be confused with percentile which indicates position in relation to a norm group.

**Grade placement at time of testing.

TABLE IX

VERTICAL LINKING OF LANGUAGE DATA

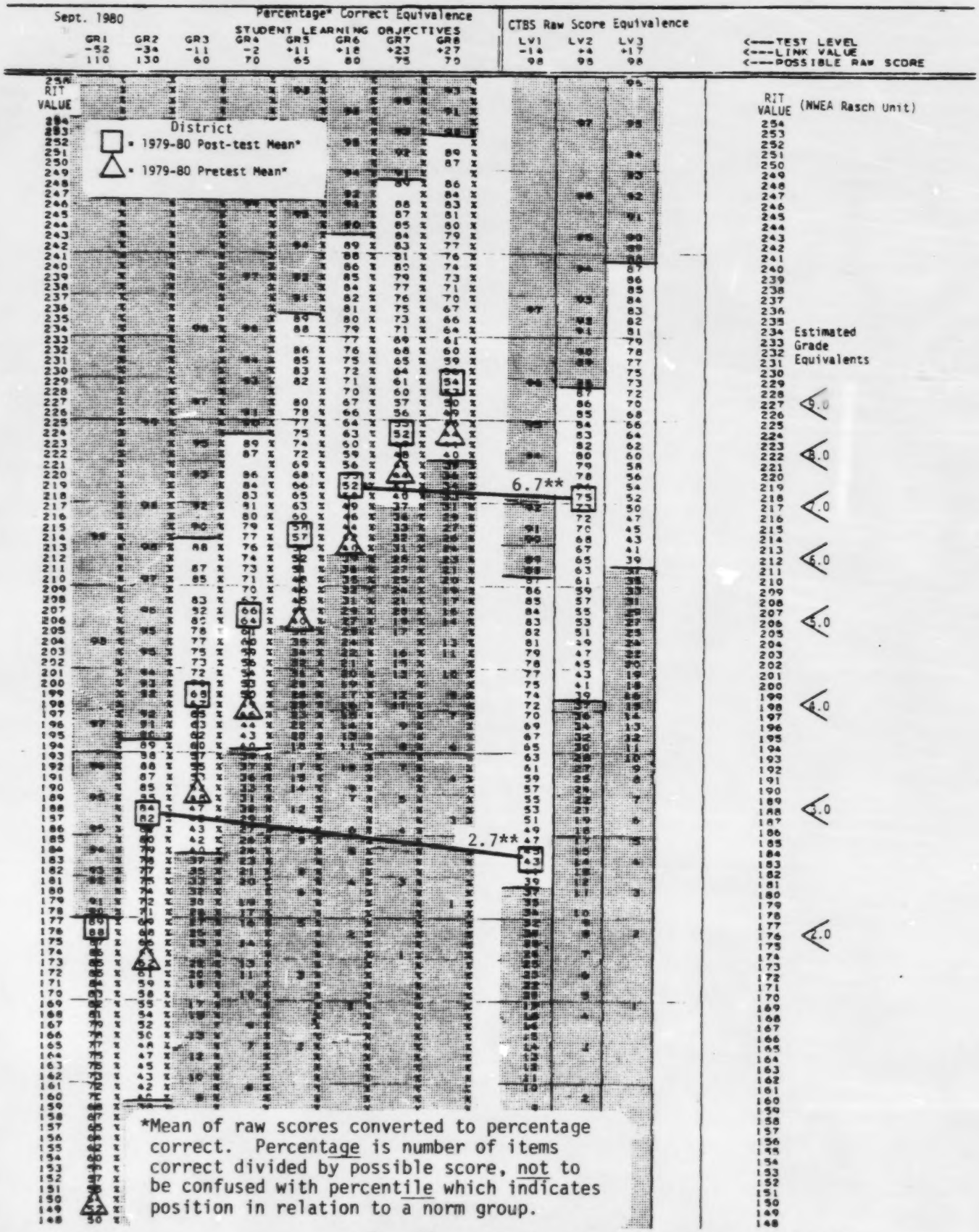


*Mean of raw scores converted to percentage correct. Percentage is number of items correct divided by possible score, not to be confused with percentile which indicates position in relation to a norm group.

**Grade placement at time of testing.

TABLE X

VERTICAL LINKING OF MATHEMATICS DATA



*Mean of raw scores converted to percentage correct. Percentage is number of items correct divided by possible score, not to be confused with percentile which indicates position in relation to a norm group.

**Grade placement at time of testing.